Running head: HERITABILITY OF A GRAMMAR CLINICAL MARKER

Grammar clinical marker yields substantial heritability for language impairments in 16-year-old twins

Philip S. Dale
University of New Mexico, Albuquerque

Mabel L. Rice
University of Kansas, Lawrence

Kaili Rimfeld
King's College London

Marianna E. Hayiou-Thomas
University of York

Author Note

Philip S. Dale, Department of Speech & Hearing Sciences, University of New Mexico; Mabel L. Rice, Department of Speech, Language, Hearing, University of Kansas; Kaili Rimfeld, Social, Genetic & Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College London; Marianna E. Hayiou-Thomas, Department of Psychology, University of York.

We gratefully acknowledge the ongoing contribution of the participants in the Twins Early Development Study (TEDS) and their families. We also appreciate advice on statistical procedures from Fruhling Rijsdijk, of the SGDP Research Centre. TEDS is supported by a programme grant (MR/M021475/1 and previously G0901245) from the U.K. Medical Research Council.

Correspondence concerning this article should be addressed to Philip S. Dale, Department of Speech & Hearing Sciences, University of New Mexico, 1700 Lomas Blvd NE Suite 1300, Albuquerque NM 87131. E-mail dalep@unm.edu

Abstract

Purpose: There is a need for well-defined language phenotypes suitable for adolescents in twin studies and other large-scale research projects. Rice and colleagues have developed a grammatical judgment measure as a clinical marker of language impairment which has an extended developmental range to adolescence.

Method: We conducted the first twin analysis, along with associated phenotypic analyses of validity, of an abridged, 20-item version of this grammatical judgment measure (GJ-20), based on telephone administration using prerecorded stimuli to 405 pairs of 16-year-olds (148 MZ and 257 DZ) drawn from the Twins Early Development Study (TEDS).

Results: The distribution of scores is markedly skewed negatively, as expected for a potential clinical marker. Low performance on GJ-20 is associated with lower maternal education, reported learning disability (age 7), and low scores on TEDS-administered language tests (16) and General Certificate of Secondary Education (GCSE) English and Maths examination performance (16). Liability threshold estimates for genetic influence on low performance on GJ-20 are substantial, ranging from 36% with a lowest 10% criterion to 74% for a lowest 5% criterion.

Conclusions: The heritability of GJ-20 scores, especially at more extreme cutoffs, along with the score distribution and association with other indicators of language impairments, provide additional evidence for the potential value of this measure as a clinical marker of SLI.

Grammar clinical marker yields substantial heritability for language impairments in 16-year-old twins

Twin studies have provided much evidence for the significant heritability of rate of language acquisition, in age levels varying from children as young as 24 months, to preschoolers, elementary school-aged children and adolescents (Spinath, Price, Dale & Plomin, 2004; Hayiou-Thomas, Dale & Plomin, 2012; Rice, Zubrick, Taylor, Gayan & Bontempo, 2014). Within this pattern of near-ubiquity of significant genetic influence on measures of language, two additional trends are notable. First, heritability for language increases with age, as it does for most cognitive measures (Hayiou-Thomas et al., 2012). Second, heritability is often higher for children who perform at lower levels of language performance (Spinath et al., 2004), suggesting stronger genetic effects for language ability below age expectations, even in samples of children screened for obvious possible etiological factors such as hearing loss or neurological disabilities. The generality of this result, however, varies with the method of ascertainment (Bishop & Hayiou-Thomas, 2008; Hayiou-Thomas, Dale & Plomin, 2014). Much higher heritabilities have been found when the samples were characterized by parental or clinical concern than when it was ascertained by low performance on psychometric measures. Bishop & Hayiou-Thomas (2008) examined the information provided by parents and concluded that the most common parental concern not reflected in the psychometric measures was difficulty in speech, which is outside the scope of this paper. Within the domain of language, they noted it appeared to be difficulties in grammar that were most likely to lead to parental concern. These two generalizations – higher heritability with impairment, and with increasing age - led us to pursue further the etiology of low grammatical ability in adolescents.

One way to further investigate the etiology of unexplained low language levels of some children relative to their age group is to focus on the language traits that are likely to be most difficult for them. Conventional language assessments are designed to examine variance across children of the same age level, and across the full range of children's abilities; an implicit goal of these measures is to enable identification of children who are high achieving, as well as those in the mid or low range of performance. An alternative approach is to investigate probable clinical markers of language impairment, using tasks that have high levels of sensitivity and specificity for identification of affected children, where sensitivity is the proportion of correctly identified children with language impairments and specificity is the proportion of correctly identified children without language impairments. Concurrent validity in the rest of the distribution is less important for a clinical marker. The primary challenge for this approach has been the development of a clinical marker that will be psychometrically robust and linguistically interpretable across a wide age range, feasible for administration to sizable samples, and therefore suitable as a phenotype for behavior genetic designs such as twin studies as well as other large-scale research studies. The psychometric properties of the phenotype are key elements of a twin investigation, which requires high levels of reliability and validity for accurate estimates of heritability, which in turn inform our understanding of possible genetic influences on variation across participants. Thus, future research on possible causal pathways of language impairments will require careful consideration of the language phenotypes, precise calculation of heritability estimates, and well-documented samples of twin children.

Grammatical judgment of finiteness marking as a clinical marker of language impairment

Research in children with Specific Language Impairment (SLI) has led to the development of several possible phenotypes suitable for twin studies. (The diversity of criteria is

also reflected in a highly diverse terminology. Bishop, Snowling, Thompson & Greenhaigh (2017) have suggested Developmental Language Disorder as a broad term with relative consensus as to criteria; however, we have chosen to continue with SLI here for continuity with existing literature.) SLI has been defined as "A language disorder that delays the mastery of language skills in children who have no hearing loss or other developmental delays." (U.S. Department of Health & Human Services. National Institute on Deafness and Other Communication Disorders (NIDCD). Specific language impairment. Available at https://www.nidcd.nih.gov/health/specific-language-impairment. Accessed August 8, 2016). Rice and colleagues (Rice, Hoffman & Wexler, 2009) addressed the need for clinical grammar markers for children with SLI older than 8 years. Previously they demonstrated that children with SLI in the age range of 3 to 8 years made systematic errors in clause construction of a particular kind: they were likely to omit grammatical morphemes needed to mark finiteness (tense, person) on verbs. In English and other languages, finiteness is required for most well-formed clauses. In English, it is marked by the morphemes for past tense (-ed and irregular past tense), third person singular –s, and specific forms of auxiliary and copula BE and auxiliary DO in site-specific locations in clause structures (see Quirk, Greenbaum, Leech & Svartik, 1985). Children with SLI omitted finiteness markers in a manner similar to younger, unaffected children but the children with SLI persisted in this immature grammar for years longer than expected (Rice & Wexler 1996; Rice, Wexler & Cleave, 1995; Rice, Wexler & Hershberger, 1998; Rice, Wexler, Wexler & Redmond, 1999). The theoretical framework for this predicted pattern is called the Extended Optional Infinitive (EOI) account. This error is seen in judgment tasks as well as productions: experimental judgment tasks confirmed that children's judgments of simple declarative clauses with or without omitted finiteness markers mirrored the grammatical patterns

found in their sentence productions. Omission of finiteness markers in obligatory contexts, in production or judgment tasks, differentiated children with SLI from unaffected children of the same age. This led to the development of the Rice/Wexler Test of Early Grammatical Impairment (TEGI; 2001), for ages 3-9 years, with high levels of sensitivity and specificity (validated independently by Spaulding, Plante, & Farinella, 2006).

For children older than 9 years, a finiteness marker is needed that is developmentally more advanced but linguistically related to the simple affirmative clauses of young children. Extending the logic of the EOI account, Rice et al. (2009) proposed that simple affirmative questions were appropriate, because (following standard linguistic theory) they require the same clause structure as declarative clauses with the addition of movement or insertion requirements. Auxiliary DO must be inserted in Wh or Yes/No questions with lexical main verbs to carry finiteness marking, as in "Where *does* the girl live?" Copula or auxiliary forms of BE must move from the base position to precede the subject, as in "He *is* running home" / "*Is* he running home?" Children who are likely to omit finiteness markers in simple declarative clauses, or accept as grammatical these clauses with omitted finiteness markers, should also be likely to apply this immature grammar in judgment tasks of questions with and without omitted DO or BE. Longitudinal growth data revealed that children with SLI persisted in lower levels of performance on this task throughout the age range of 6 to 15 years, compared to unaffected children who were at consistently high levels of performance throughout this age range.

Fine-grained linguistic analyses of children's utterances support the notion that children with SLI, although prone to omission of finiteness markers, nevertheless control the syntactic structures needed to generate clauses and avoid many possible kinds of grammatical errors. Their tendency to omit auxiliary DO in questions, for example, does not affect their high levels

of productivity and accuracy with main verb forms of DO, as in “I do my homework at night.” (Rice & Blossom, 2013). Children with SLI are likely to omit third person singular –s in a simple imitation task but nevertheless avoid otherwise ungrammatical clauses in their inaccurate imitations (Abel, Rice & Bontempo, 2015). For example, when they hear “the girl gives her doll to a friend” they are likely to say ”the girl give her doll to a friend” but not “the girls gives the doll to a friend” or “the girl gives the doll.” Such evidence highlights the distinctiveness of finiteness marking as a clinical marker, and the ways in which children with SLI demonstrate strengths as well as weaknesses in grammar.

The phenomenon of omitted finiteness marking by typical children and for an extended time by children with SLI has been further explored empirically to evaluate its validity, as well as to compare it with other aspects of language impairment. Bishop, Adams & Norbury (2006) evaluated finiteness markers as a phenotype in a sample of 6-year-old children recruited from the Twins Early Development Study (TEDS), as well as a non-word repetition phenotype. Using DeFries-Fulker analysis to analyze etiology of low performance, they found significant heritability for both phenotypes (.74 for finiteness marking; .61 for nonword repetition) but largely non-overlapping mechanisms of inheritance, with distinct genetic origins (genetic correlation r_g = .09), suggesting that finiteness marking is qualitatively different from non-word repetition. The extent to which the grammatical property of finiteness-marking is associated with other putative predictors of the rate of children’s language acquisition also suggests that it may be a distinct construct. Consistently across studies, several predictors have been identified for a range of language outcomes; they include performance on nonverbal intelligence assessments, single word vocabulary assessments, and mother’s education (as a surrogate index of home influences). However, across multiple studies (Rice, Hoffman & Wexler, 2009), these

variables do not predict the growth of finiteness markers, suggesting distinct etiological sources for the markers.

Finiteness and non-word repetition are not the only dimensions of language that show higher rates of differences in developmental growth trajectories in children with SLI. Rice & Hoffman (2015) document that children with SLI trail their unaffected age peers in receptive vocabulary acquisition throughout childhood, leveling off at persistently lower levels in adolescence and early adulthood. Although the protracted lower vocabulary acquisition of the affected group is noteworthy and of clinical concern, the gap between the affected and unaffected groups do not appear to be quite as large for vocabulary acquisition as in the studies of the finiteness marker. Further, as reported by Spaulding et al. (2006), tests of single word vocabulary generally show unacceptably low levels of sensitivity and specificity for detection of children with SLI.

Growth modeling studies of finiteness markers and receptive vocabulary of children with and without SLI (see Rice, 2002, 2013 for summaries) have determined that the growth parameters do not differ substantially between groups, whereas the intercept, or starting point, does. It is as if the start-up of the growth trajectory is delayed for the affected group, but when growth begins the growth trajectories are robust for both groups. The catch is that in the pre-adolescent period growth begins to decelerate for both groups of children, leaving the children with SLI at a lower level that persists into adolescence and beyond. Rice argues that these results are consistent with the view that the etiology of SLI involves an age-related growth signaling dysfunction – a delay, along with other possible changes - in the underlying genetic mechanisms contributing to language acquisition, a dysfunction that does not establish necessary

cortical neuronal infrastructures at the right times for children with SLI to match the growth trajectory outcomes of unaffected children.

The present study

The present study is the first twin investigation of a grammar clinical marker in adolescence previously validated in studies of children with Specific Language Impairment (SLI). The study has been conducted with a non-clinically selected sample of children, in which language impairments are identified by low performance on measures of several aspects of language directly administered to participants. For comparison, we also have parental reports of learning disabilities and/or dyslexia. This sample allows us to avoid the well-known biases of clinical referral, and to have uniform measures on all subjects. The twins in the present sample are 16-year-olds, allowing for investigation of the heritability of limited grammatical ability in an age range where heritability estimates may be higher than at younger ages and when mastery of the adult grammar is expected for the great majority of participants. We address several phenotypic questions concerning the distribution and validity of the measure before turning to the primary question of the paper, the etiology of variation in this measure (#4):

1. What is the distribution of scores on the proposed clinical marker, and their internal consistency, in a non-clinically selected sample with a full range of language ability?
2. How well does variance in this measure relate phenotypically to other language measures and maternal measures at this age in terms of correlation across the full range of variability?
3. How well does low performance on an abridged grammatical judgment task relate to other indicators of low language and literacy achievement?
4. What is the etiology of placement at the low extreme on GJ-20, and how does the etiology change with increasingly strict criteria for low performance?

Method

Participants

The broad sampling frame for the present study was the Twins Early Development Study (TEDS), a longitudinal study of twins born in England and Wales in 1994, 1995, and 1996 (Oliver & Plomin, 2007; Haworth, Davis & Plomin, 2012). After checking for infant mortality, all families identified by the UK Office of National Statistics (ONS) as having twins born in those years were invited to participate in TEDS when the twins were about 18 months of age. The twins were assessed at 2, 3, and 4 years of age using parent questionnaires which included measures of language, cognitive, and behavioral development. They have continued to be assessed in these and other domains with a variety of methods including telephone assessment, parent-administered tests, teacher National Curriculum ratings, and increasingly from age 10 years on, web-based assessment.

A subset of TEDS twins was selected on the basis of the parental assessment at 4 years for an in-depth, in-home assessment at 4.5 years (Kovas, Hayiou-Thomas, Oliver, Dale, Bishop & Plomin, 2005; Hayiou-Thomas, Kovas, Harlaar, Plomin, Bishop & Dale, 2006). They were assessed on an extensive battery of language and nonverbal measures. Twin pairs were excluded if either member had any major medical or perinatal problems, documented hearing loss, or brain damage. Maternal education levels were comparable to the overall TEDS sample, as well as to UK ONS census data (Kovas, Haworth, Dale & Plomin, 2007). Only participants for whom English was the first language spoken at home were selected. A total of 834 twin pairs provided at least some data, with 787 pairs providing complete in-home assessment data.

Because one purpose of the in-home study was to examine the etiology of language and cognitive impairments, the sample was constructed to over-represent children whose performance was low. For this purpose, the age 4 parent report of vocabulary (lowest 5% on a checklist) and grammar (child reported as not yet talking, talking in one-word utterances, or talking in 2 or 3 word utterances), and expressions of concern over speech and language development (selecting "his/her language is developing slowly") were used (see Dale, Price, Bishop & Plomin, 2003 for more information about the age 4 measure). About 60% of the sample was identified in this way; the remainder was a random sample of twin pairs who did not meet the criteria for low performance (the 'control sample'). As documented in Kovas et al. (2005), the resulting combined sample yielded distributions on all measures which were unimodal and in most cases near-normal, with means that were .5 to 1.0 SD below the mean for the control sample. For all further analyses in the present paper, the combined sample was used.

In order to facilitate future longitudinal analysis, the target sampling frame for the present study were the 834 twin pairs and parents who had provided at least some data for the 4.5 year study. Of these, 153 families had withdrawn from TEDS, were classified as medical exclusions (e.g., later diagnosis of autism spectrum disorder and other genetic disorders), were living outside the UK, or had unusual family circumstances. This resulted in 681 families selected for contact. Of these, 440 gave initial consent, and with some further loss due to scheduling and technical problems, 405 twin pairs provided complete data from both twins. Participants were recruited and tested as described below, for a total sample of 405 pairs (148 MZ and 257 DZ; mean age 16.6 yrs, SD=.82).

Measures and Procedure

Grammatical judgment of finiteness marking task. This task was an abridged version of the one studied by Rice et al. (2009). That measure contains 40 items, half affirmative *wh*-questions and half yes/no questions. Each of these 20-item categories is further divided into 10 items with BE/DO forms present or omitted; and half of these subsets are grammatical and half ungrammatical due to omitted BE or DO forms. All items are affirmative, and the argument structures and semantics are familiar and simple. Correlations calculated on the data of Rice et al. (2009), found high levels of association between the BE and DO items (Pearson $r = .74$, $p < .001$, N = 143). There is a high association of performance on the DO items with the full 40-item test (Pearson $r = .92$, $p < .0001$, N = 143). Consequently, it was judged that the 20 *wh*-questions by themselves captured the essential variability between children with SLI and those without. Therefore, in the interests of brevity for telephone administration, those items alone constituted the present task, labelled GJ-20. A few minor wording changes were made for appropriateness in UK English, e.g., substituting 'chips' for 'French fries', and changing the response choices from 'right' vs. 'not so good' to 'right' vs. 'not quite right'. Items were presented in the same, randomized order for all participants, following 10 practice items. Two example test pairs, each with one grammatical sentence and one not, are 'Where does the bug like to sleep?/When _ you like to sleep?' and 'What does she like to drink'/What _ you like to drink?" Note that the underlying construct to be evaluated is the obligatory context for the presence of auxiliary DO, across variations in subjects and *wh*-forms. Any possible effect of variability in subject form (2nd vs. 3rd person singular) or *wh*-form (*what, where, when, why*) variability would reduce the accuracy of the assessment, which is designed to assess the underlying abstract requirement of *do* across multiple grammatical contexts.

The test was administered to each participant over the telephone, typically requiring less than five minutes. The stimuli were pre-recorded with a female voice (a native speaker of UK English with a standard, 'BBC English' dialect) and the 24 bit WAV files were played through the PC to standardise administration. Each test item was saved as an individual file, to give the administrator control over presentation and to allow items to be replayed if needed. The ComPack universal telephone audio interface included a direct connection to a cellular telephone. (More complete information on recording and playback is available from the first author on request.) Each twin was tested individually, on some occasions one after the other and sometimes on different days. Participants were asked to move to a quiet space in their homes and away from their twin. Volume and clarity of phone line was also checked to make sure the participant had no problems hearing the test. A short set of instructions were given, explaining that a series of sentences would be played over the telephone and that the participant's task is to listen carefully to each sentence and say whether it sounded 'right' or 'not quite right'. The participant was then asked if they understood the instructions before 10 practice items were administered. Feedback was given for these practice items and correct response was given if the participant got a question wrong. The administrator then proceeded to play the 20 test items where no feedback was provided. Participants' responses (binary, as described above) were recorded simultaneously to administration, and were scored online on data sheets.

In this study, the stimuli were delivered by telephone, likely with some decrease in acoustic quality due to reduced bandwidth, which might especially affect certain phonemes such as the fricative /s/. However, the critical elements of the stimuli, for which the task was in part to detect their presence or absence, were forms of BE and DO, which are full syllables or more. As noted above, the volume and clarity of the phone connection was checked with participants

before proceeding, and the participant had the option to have items repeated. Furthermore, if these elements were in fact more difficult to hear over the phone, the effect would be to reduce discrimination between grammatical and ungrammatical forms, reducing the ability to discriminate the two categories of stimuli and thus lowering both the reliability of the measure and mean level of performance. In summary, these issues render the present results, including estimates of heritability, very conservative and likely underestimates.

Test-retest reliability for this measure can be computed from the data collected by Rice et al. (2009). We extracted the 20 items used in the present study from the full 40-item list, and calculated reliability of A' across a one year retest interval. For age 14-15 (n =179), r = .74; for age 15-16 (n = 135), r = .71; and for age 16-17 (n = 112), r = .74.

Age 16 language measures from the larger TEDS study. Two language measures had been administered by internet to the first two of the four birth cohorts in the larger TEDS study. The first was Section B (the multiple choice portion) of Form 1 Senior Version of the Mill Hill Vocabulary Scale (Raven, Raven & Court, 2010). For individuals at this developmental level, the test includes 33 words, beginning with 'fascinated' and ending with 'minatory'. Internal consistency of alpha = .81 and test-retest correlation of *r* = .64 are reported for this test. The second was the Figurative Language subtest of the Test of Language Competence – Expanded Edition (Wiig, Secord & Sebers, 1989). This subtest assesses the interpretation of idioms and metaphors, which require both rich semantic representation and an awareness of the ambiguity of many expressions between their literal and figurative meanings. The participant hears a sentence orally and chooses one of four answers, presented in both written and oral forms. Internal consistency of alpha = .69 and test-retest correlation of r = .71 are reported for this test. No grammar measure was included at 16, due to the lack of a well-validated receptive measure for

grammar at this age which was sensitive to individual differences across the full range. In addition, previous work in TEDS (Dale, Harlaar & Plomin, 2010) has generally found substantial phenotypic and genetic correlations between vocabulary and grammar measures. (These correlations between vocabulary and grammar are not inconsistent with the focus on the grammatical feature of finiteness marking as a marker of SLI, because they are measures of the relationship across the full distribution rather than at the low extreme, and also because the grammatical measures were substantially broader than finiteness marking.) Because these tests were administered to only two of the four birth cohorts in the TEDS sample, the Vocabulary measure was available for only 151 pairs (37.3%), and the Figurative Language measure for 146 pairs (36.0%) of those participating in the present study.

Additional educational measures for students. Based on questionnaires completed by parents when their children were 7 years old, we identified children reported as having learning difficulties ('Does either of the twins have difficulties with their learning?") and dyslexia specifically ('If YES, what is the difficulty that each twin has?" with 'Difficulties in learning to read/dyslexia' as an option selected).

At age 16, we obtained examination results in English and Mathematics as part of the UK nationwide examination for educational achievement at the end of compulsory education, the General Certificate of Secondary Education (GCSE). GCSE examinations are held in a diverse range of academic areas, but English, Mathematics, and Science are compulsory. The exams cover courses which begin around age 14 and are completed around age 16. GCSEs are graded from A*, the highest grade, to G, the lowest passing grade (no information about failing results was available). There is no mandatory number of GCSEs to be taken, but students commonly take between 8-10 subjects, and receiving five or more grades in the range of A*-C is typically a

requirement for going on to further education. Shortly after completion of the GCSEs, each TEDS family was sent a results form, followed as necessary by telephone reminders. Scores for English (a composite of English language and English literature) and Mathematics were standardized, based on the larger TEDS sample, for the present analysis. Further details of the examination, data collection and scoring, as well as the high accuracy of parent reports of GCSE scores, are available in Shakeshaft et al. (2013).

Additional family measures. At entry in the study, mothers provided information on their educational attainment ('qualifications'). These were scored on an 8-point basis, ranging from 0 = none through 4 = A level exams (taken at age 18 by students anticipating university education) to 7 = undergraduate degree and 8 = postgraduate degree.

At twin age 9, parents provided information about family history of early language and/or reading learning difficulties. The family history variable was coded as 1 if any first degree relative (mother, father, older brother, older sister) was reported as having either type of learning difficulty, and 0 otherwise.

Analysis

Following Rice, Wexler and Redmond (1999), the primary dependent variable for the GJ-20 task was $A' = 0.5 + (y-x)(1+y-x) / 4y(1-x)$, where x = proportion of false alarms (saying 'yes' to an ungrammatical item) and y = hits (saying 'yes' to a grammatical item). A', which has the range 0 – 1.00, is more appropriate than percent correct because it adjusts for children's tendency to default to 'yes' responses, and the ungrammatical items on the test might reasonably be expected to yield 'yes' responses from affected children (see also Linebarger, Schwartz & Saffran, 1983).

Phenotypic analyses of this measure included examination of the GJ-20 score distribution and internal consistency (Question #1). This was followed by an examination of the validity of GJ-20 scores as measured by correlations with the demographic variables of gender and maternal education as a measure of SES, and with the two standardized language measures, Vocabulary and Figurative Language, at two levels (Questions #2 and #3). The first was an overall correlation or other measure of association across the entire distribution of individual differences. The second was an examination of the agreement in classification as low performance, using a criterion of lowest 10% for the two measures being compared. This criterion is often used as a cutoff in research and clinical work (Dale et al., 2003; Reilly, Wake, Ukoumunne, Bavin, Prior, Cini, Conway, Eadie & Bretherton, 2009; Reilly, Wake, Ukoumunne, Bavin, Prior, Cini, Conway, Eadie & Bretherton, 2010; Archibald & Joanisse, 2009). Phenotypic analyses were based on one randomly selected twin from each pair, to preserve independence of data, with one exception. For the analyses exploring whether low GJ-20 scores were associated with low scores on other language and academic measures (Research question 3), where statistical power was an issue, alternative approaches were used which allowed the use of data from both twins in each pair with appropriate adjustment, as explained below.

For the etiological analyses (Question #4), all measures were standardized for the entire sample, and scores were corrected for the linear effects of age and sex, as these can inflate twin similarity. To maximize power, both same-sex and opposite-sex DZ twins were included; we note that sex differences in etiology have generally not been found for language measures in TEDS data (Kovas, Haworth, Dale & Plomin, 2007; Shakeshaft et al., 2013). Because the GJ-20 measure was devised as a clinical marker, the etiological analyses of the measure were conducted only for placement in the low performance group. We defined probands as

adolescents who scored in the lowest 10% of the sample. We calculated probandwise concordance as the ratio of the number of probands in concordant pairs to the total number of probands. Greater concordance for MZ pairs than for DZ pairs suggest genetic influences. Due to the small sample size, and the unknown significance of variation within the normal range on this measure, for the main etiological analysis we used liability threshold modeling, which is based on dichotomizing the measures, and then comparing the degree of agreement for MZ twin pairs, who are genetically identical, with the agreement for DZ pairs who share 50% of the segregating genes on average. We also explored the etiology of cutoffs other than lowest 10%. This exploration is not an evaluation of specific clinical criteria for identifying language impairment; it is simply a way to determine the effect of varying the stringency of cutoff.

Finally, we note that the current analyses do not include multivariate etiological analyses of the new measure with the two psychometric language measures, as the sample size for individuals with all the relevant measures does not provide sufficient statistical power for interpretable results.

Results

Representativeness of the sample

The representativeness of the present sample relative to the full and relatively population-representation TEDS sample can be evaluated with respect to maternal education (as a measure of SES) and two language measures administered as part of the main age 16 battery: Vocabulary and Figurative Language. As noted earlier, these language measures were obtained for only a portion of the present sample; however, as that was based on birthdate alone it can be considered an unbiased sample. Utilizing the 0-8 scale used in previous TEDS publications to measure

maternal education at entry to the study, the present sample had non-significantly higher maternal education than remainder of the TEDS sample participating at age 16 (4.14 vs. 3.96, t(6763)=1.66, p=0.097, d = .09). As expected, given the selection of the original sample at age 4.5 years, the present sample had significantly lower vocabulary scores than the remainder of the TEDS sample, though the difference was small (14.76 vs. 15.47, t(2592)=1.97, p=0.049, d = .17). The difference between the groups on figurative language was not significant (10.44 vs. 10.27, t(2462)=0.778, p=0.437, d = .07).

Phenotypic analysis of the distribution and internal consistency of the GJ-20 measure (Question #1)

Except as noted, the phenotypic results presented here are based on the selection of one randomly selected twin from each pair, to preserve independence of data. Figure 1 presents a histogram of GJ-20 scores (A', as defined above). As desired for a clinical marker, the score distribution has substantial negative skew of -3.25 (for comparison the skew for vocabulary and figurative language measures are .39 and -.58, respectively). Because GJ-20 was marginally correlated (r = .10, p = .08) with age, an age-corrected score was computed for use in all analyses. Vocabulary and Figurative Language were also age-corrected. Table 1 summarizes the overall distribution of age-corrected GJ-20 scores, by sex and zygosity, along with that for Vocabulary and Figurative Language. The GJ-20 score did not differ significantly by gender, nor did either of the other measures (all p > .05). GJ-20 did not differ significantly by zygosity. For both Vocabulary and Figurative Language, DZ twins scored significantly higher (for Vocabulary, t[149] = 2.47, p = .02; for Figurative Language, t[144] = 2.61, p = .01), but the effect size is trivial, representing less than 3% of the variance.

Conventional measures of internal consistency such as Cronbach's alpha are not appropriate for A' as it is a function of differential performance on two subsets of the items (grammatical and ungrammatical). Only a subset of the possible divisions of the items into halves would permit calculation of A', and they would be based on very few items. A total score correct, based on all 20 items, is highly correlated with A' (r = .94), and does permit calculation of alpha, which is .575.

The validity of the GJ-20 across the full range of variability (Question #2) and as an indicator of low performance (Question #3)

Table 2 includes the intercorrelations among the language measures, as well as with maternal education. The age-adjusted GJ-20 scores were not correlated significantly with Vocabulary (.12) but were correlated significantly, but weakly, with Figurative Language (.24). All three measures were correlated significantly with maternal education, though the correlations were low (.16-.24).

Given the clinical focus of the GJ-20 measure, the most relevant phenotypic analyses are based on placement in the low extreme range, and its relation to demographic, language, and other measures. A criterion of lowest 10% was adopted for each of the three language measures, for consistency with each other and with a commonly used clinical cutoff.

Table 3 summarizes analyses of the validity of low GJ-20 scores as an indicator of language and literacy difficulties. It also includes the result that there was not a sex difference in GJ-20 score. Because the overall sample size is not large, and especially small for the low GJ-20 group, obtaining adequate statistical power for comparisons of the low and normal GJ-20 is essential. For this reason, we adopted a more complex analysis for the continuous comparison measures (maternal education, Age 16 Vocabulary and Figurative Language, GCSE English and

Mathematics), which allowed us to use the data from all twins. We fit structural equation models to the variables of interest (e.g. vocabulary score in twin 1 and 2) in 4 sub-sets of data, according to the twin-pair scores on the GJ-20 measure: where both twins are above threshold and therefore in the normal range (NN), both below (LL) and the discordant pairs (NL and LN). Depending on the subgroup, the covariance models are specified by means of a standard deviation for the normal group and/or one for the low group (sdN, sdL). Similarly, the Means models are specified by a parameter for the normal and/or one for the low group (mN, mL). We specified one overall twin correlation to account for non-independence of observations. A model representing the null hypothesis is then tested by equating mN and mL. Significance is evaluated by likelihood ratio (Chi-square) test (Miles, 2003).

Based on these analyses, twins with GJ-20 scores in the lowest 10% had mothers with significantly lower education, scored significantly lower on TEDS web–administered tests of vocabulary and figurative language at age 16, and had significantly lower GCSE scores in English and Mathematics. There is no comparable analysis for non-independent categorical data; we have therefore conducted analyses utilizing just one randomly selected twin from each pair (over-conservative, as it discards half of the data) and also utilizing both twins (over-liberal, as it assumes independence of data). The true significance level lies somewhere between the two values. The two approaches both yield significant differences between participants scoring in the lowest 10% on the GJ-20 and those above this threshold, for frequency of reported learning disabilities; there were nonsignificant differences for reported dyslexia, and for family history of language/learning disabilities.

Etiological (genetic) analysis (Question #4)

The probandwise concordance for placement in the low score category on GJ-20 is shown in Table 4 for the 10% criterion as well as 7% and 5%. The figures are higher for MZ twins than for DZ twins, suggesting genetic influence, and this trend appears to increase with increasingly strict criteria.

Liability threshold analyses to estimate the genetic and environmental influences for the GJ-20 measure were conducted for same three criteria: lowest 10%, 7%, and 5%; they are summarized in Table 5. (A represents genetic influence as % of variance in the measure; C is shared environmental influence; and E is non-shared environmental influence. E also includes measurement error.) Consistent with the probandwise concordances shown in Table 5, genetic influence on the grammatical judgment task is significant, and the point-estimates suggest a substantial increase in heritability with more stringent cut-off criteria. rising from .36 to .74. Given the wide confidence intervals provided by liability threshold analyses for a sample of this size, however, the differences among the three estimates were not significant.

Discussion

The present study is the first examination of the grammatical judgment task in a non-clinically selected sample of adolescents, which while not fully population-representative, did include the full range of language ability. One goal of the project was to determine whether a judgment task for finiteness marking on English verbs can provide evidence in a non-clinically selected sample of language deficit, such that it might be used as a clinical marker. A second major goal was to determine the etiology of this measure. A related goal was to explore whether heritability would increase with more stringent cut-offs, as has been suggested for other measures of language impairment in the literature. The first three research questions focused on

the variation in scores on the abridged grammatical judgment task: the shape of the distribution and the reliability of the measure (Research question 1), and correlations with other language measures (Research question 2) and with measures of language and literacy difficulties (Research question 3). The distribution of GJ-20 scores is highly skewed, as is desirable for a clinical marker; this suggests that variation in the low extreme is more informative than in the remainder of the score range. GJ-20 scores are weakly correlated with other language measures over the entire distribution (Table 2; r = .12 and .24), and they are significantly and more strongly associated with reported learning difficulties and with lower scores on age 16 Vocabulary, Figurative Language, and GCSE English and Mathematics (Table 3, *d* = .76, .72, .67, and .62 respectively, which correspond to correlations in the range of .30-.35) .

With respect to the fourth and primary research question, concerning the etiology of low performance on the GJ-20, liability threshold modeling confirms substantial genetic influence, and the estimates rise with increasingly stringent cutoffs (Table 5) although the wide confidence intervals do not let us document a significant trend. We can only conclude that heritability is substantial at all three cutoffs, and at 5% it merits the term ‘high’. This result is notable because it is an exception to the pattern described at the beginning of this paper, in which heritability estimates are generally lower for impairment defined by low test scores (as in the present study) relative to those found on the basis of parental or clinical concern. It is also notable that the heritability estimates here for impairment on GJ-20 defined by cutoffs equal to the lowest 7% and 5% (.47 and .74) are comparable to the .74 estimate derived by Bishop, Adams and Norbury (2006) for an elicited production test of finiteness markers in 6 year olds, although in that case the high heritability was obtained at a more liberal cut-off (13%) than was used here. Bishop et al. did not include longitudinal analyses which would evaluate the stability of poor performance

on finiteness markers, and therefore the similarity of heritabilities at 6 and 16 can only be taken as suggestive.

High heritability is not in itself either a necessary or sufficient condition for a proposed clinical marker, and certainly an analysis of the etiology of performance on the grammatical judgment test has its own interest. However, we propose that the combination of a theoretical rationale for the importance of finiteness marking, well-documented evidence for the heritability of language impairment, and the clear evidence for substantial heritability of the GJ-20 measure constitutes a 'mutually supportive' body of evidence for the relevance of the marker for language impairment. If this conclusion is supported by further research, it suggests that a grammatical judgment task focused on finiteness such as GJ-20 might be useful for screening, and perhaps identify, residual, unidentified language impairments in older children and adolescents (Catts, Fey, Weismer & Bridges, 2014; Nation, Clarke, Marshall & Durand, 2004).

Limitations and Future Directions

Although it is a strength of this study that it examined the performance of a non-clinically selected sample which included the full range of language ability, this meant that direct indicators of clinically defined language impairment were not available, only parent-reported learning difficulties and dyslexia. We do not have any evidence that would allow us to evaluate the accuracy and completeness of these parent-reported difficulties. Thus the associations reported in Table 3 can only be taken as approximations. A second limitation concerns sample size. Although the present sample is the largest ever tested with GJ-20 or similar measures, it is relatively small by behavior genetic standards. A consequence of the limited sample size and the use of liability threshold modeling is that the confidence intervals are extremely wide, which

limits the interpretation that can be placed on any specific value, and rules out for now a firm conclusion concerning changes in heritability with increasingly strict criteria. Further, the samples size does not have enough power to justify multivariate analyses, which would determine how much of the relationship between GJ-20 and other language measures or indicators of deficit is due to genetic or environmental influences that affect the relevant pair of measures. The modest internal consistency of the GJ-20 measure (alpha = .575) is also a limitation. It is hoped that the evidence from this study will lead to the inclusion of this relatively brief measure or ones similar to it in future studies of language development and impairment. The GJ-20 or similar tasks have considerable promise for use in research as a phenotype for investigation of causal pathways of a grammar weakness across a wide age range of participants.

The present study did not include measures of working memory, particularly non-word repetition, which has also been proposed as a clinical marker (or 'endophenotype') for Specific Language Impairment (Bishop, North & Donlan, 1996; Dollaghan & Campbell, 1998). Few studies have included both. Archibald and Joanisse (2009) examined several language and memory measures for a group of 400 6-9 year olds, and concluded that sensitivity and specificity were better for sentence recall, taken as a measure of grammatical ability, than for non-word repetition. More research is needed here, particularly to determine the extent to which these measures vary in their validity as assessments of the same impairment, or whether they have differential sensitivity to diverse impairments.

The participants in the study were not screened for hearing loss, although currently it has been estimated that 10-20% of teenagers show some sign of hearing loss, either congenitally or due to excessive exposure to loud music and other sounds (Su & Chan, 2017). It is unknown whether the present sample, who had volunteered their participation in a telephone-based study,

included adolescents who had a loss which could affect performance. Individuals with substantial loss might simply have not volunteered. Furthermore, as noted above, participants were checked for volume and clarity of the stimuli, and had an opportunity for repetition. For these reasons, we do not believe that hearing loss played a significant role in these results.

The substantial heritability of the GJ-20 measure suggests that it might be a good candidate for molecular genetic research, such as Genome-Wide Association Studies (GWAS; McCarthy et al., 2008; Paracchini, 2011). However, there is much converging evidence that the effect of individual genes (or SNPs) is very small, and therefore very large samples (thousands, if not tens of thousands) are needed for this work. In addition, replication of positive GWAS findings has been scarce. Harlaar et al. (2014), for example, analyzed performance on four web-based measures of receptive language in 2,329 12-year-olds. None of the associations met the usual requirement of genome-wide statistical significance that corrects for multiple testing of more a million potential associations, and the strongest association found did not replicate in an additional sample of 2,639 12-year-olds. Thus very large samples, and probably new genetic and statistical techniques such as complex polygenic scores (Krapohl et al., under review; Dudbridge, 2013) are likely to be needed for successful ‘gene-hunting’ (in the broadest sense of prediction from DNA to behavior) to be successful. Precision of measurement of the language phenotype also influences the power of molecular genetic research designs; thus the importance of continuing research on the most valid clinical markers of language impairment.

Conclusion

We have examined the sensitivity of a brief grammatical judgment task, the GJ-20, to variation in the language ability of adolescents, both across the full range of ability, and with respect to

indicators of impairment. GJ-20 scores have several properties which suggest clinical utility, including substantial negative skew, association with parent reported learning difficulties, and heritability in the moderate-to-high range depending on the criterion used for low performance. The results suggest a continuity of impairment from the more intensively studied grammatical difficulties of kindergarten and elementary school children to those observable in mid-adolescence, characterized by a persistent weakness in the requirement for finiteness marking in obligatory syntactic sites. Overall, the GJ-20 shows considerable promise for both clinical and research-based assessment.

References

Abel, A.D., Rice, M.L. & Bontempo, D.E. (2015) Effects of verb familiarity on finiteness marking in children with SLI. *Journal of Speech, Language, and Hearing Research*, 58:360-372.

Archibald, L., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52*, 899.

Bishop, D.V.M., Adams, C. V., & Norbury, C.F. (2006). Distinct genetic influences on grammar and phonological short-term memory deficits: Evidence from 6-year-old twins. *Genes, Brain, and Behavior, 5*, 158-169.

Bishop, D. V. M, & Hayiou-Thomas, M. E. (2008). Heritability of specific language impairment depends on diagnostic criteria. *Genes, Brain & Behavior, 7*, 365-372.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhaigh, T. (2017). Phase 2 of catalise: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*. Doi: http://dx.doi.org.libproxy.unm.edu/10.1111/jcpp.12721.

Catts, H. W., Fey, M. E., Weismer, S. E., & Bridges, M. S. (2014). The relationship between language and reading abilities. In J. B. Tomblin & M. A. Nippold (Eds.), *Understanding individual differences in language development across the school years*. New York: Psychology Press, pp. 145-165.

Dale, P. S., Harlaar, N., Hayiou-Thomas, M. E., & Plomin, R. (2010). The etiology of diverse receptive language skills at 12 years. *Journal of Speech, Language, and Hearing Research, 53*, 982-992.

Dale, P. S., Price, T.S., Bishop, D.V.M., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research, 46*, 544-560.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics, 9*.

Harlaar, N., Meaburn, E. L., Hayiou-Thomas, M. E., Welcome Trust Case Control Consortium 2, Davis, O. S. P., Docherty, S., … Plomin, R. (2014). Genome-wide association study of receptive language ability of 12-year-olds. *Journal of Speech, Language, and Hearing Research, 57,* 96-105.

Haworth, C. M. A., Davis, O. S. P., & Plomin, R. (2012). Twins Early Development Study (TEDS): A genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood. *Twin Research and Human Genetics, 16*

Hayiou-Thomas, M. E., Kovas, Y., Harlaar, N., Plomin, R., Bishop, D.V.M., & Dale (2006).Common etiology for diverse language skills in 4-1/2-year-old twins. *Journal of Child Language, 33*, 339-368.

Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2012). The etiology of variation in language skills changes with development: A longitudinal twin study of language from 2 to 12 years. *Developmental Science, 15,* 233-249.

Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2014). Language impairment from 4 to 12 years: Prediction and etiology. *Journal of Speech, Language, and Hearing Research, 57,* 850-864.

Kovas, Y., Haworth, C. M. A., Dale, Philip S., & Plomin, R. (2007), The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development, 72,* Serial No. 288.

Kovas, Y., Hayiou-Thomas, M. E., Oliver, B., Dale, P. S., Bishop, D.V.M., & Plomin, R. (2005). Genetic influences in difference aspects of language development: The etiology of language skills in 4.5-year-old twins. *Child Development, 76*, 632-651.

Krapohl, E., Paten, H., Newhouse, S., Curtis, C. J., von Stumm, S., Dale, P. S., Zabaneh, D., Breen, G., O'Reilly, P. F., & Plomin, R. (under review). Multipolygenic score (MPS) models predict 11% variation in educational achievement and 5% in general cognitive ability.

Leonard, L. B. (2014). *Children with specific language impairment*. Cambridge, MA: Bradford.

Linebarger, M., Schwartz, M. & Saffran, S. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13*, 361-392.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ionnidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty, and challenges. *Nature Reviews: Genetics, 9*, 356-369.

Miles, J. (2003). A framework for power analysis using a structural equation modelling procedure. *BMC Medical Research Methodology, 3*, 27.

Nation, K., Clarke, P., Marshall, C.M., & Durand, M. (2004). Hidden language impairments in children: Parallels between poor reading comprehension and Specific Language Impairment? *Journal of Speech, Language and Hearing Research, 47*, 199-211.

Oliver, B. R., & Plomin, R. (2007). Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10*, 96-105.

Paracchini, S. (2011). Dissection of genetic associations with language-related traits in population-based cohorts. *Journal of Neurodevelopmental Disorders*

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*, New York: Longman, Inc.

Raven, J., Raven, J. C., & Court, J. H. (2010). *Mill Hill Vocabulary Scale*. Oxford: Oxford University Press.

Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., Conway, L., Eadie, P & Bretherton, L. (2009). The Early Language Study in Victoria Study (ELVS): A prospective, longitudinal study of communication skills and expressive vocabulary development at 1, 12, and 24 months. *International Journal of Speech-Language Pathology, 11*, 344-357.

Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., Conway, L., Eadie, P & Bretherton, L.(2010), Predicting language outcomes at 4 yers of age: Findings from Early Language in Victoria Study. *Pediatrics, 126*, 1530-1537.

Rice, M.L. (2012). Toward epigenetic and gene regulation models of specific language impairment: Looking for links among growth, genes, and impairments. *Journal of Neurodevelopmental Disorders, 4*, 27

Rice, M.L. (2013). Language growth and genetics of specific language impairment. *International Journal of Speech-Language Pathology, 15*, 3, 223-233.

Rice, M. L., Zubrick, S. R., Taylor, C. L., Gayan, J., & Bontempo, D. E. (2014). Late language emergence in 24-month-old twins: Heritable and increased risk for late language emergence in twins. *Journal of Speech, Language, and Hearing Research, 57*, 917-928.

Rice & Blossom, 2003, What do children with Specific Language Impairment do with multiple forms of DO? *Journal of Speech, Language, and Hearing Research, 56*, 222-235

Rice, M.L. & Hoffman, L. (2015). Predicting vocabulary growth in children with and without Specific Language Impairment (SLI): A longitudinal study from 2 ½ to 21 years of age. *Journal of Speech, Language, & Hearing Research*, 58:345-359.

Rice, M. L., Hoffman, L., and Wexler, K. (2009). Judgments of omitted BE and DO in questions as extended finiteness markers of Specific Language Impairment (SLI) to 15 years: A study of growth and asymptote. *Journal of Speech, Language, and Hearing Research, 52*, 1417-1433.

Rice, M. L., Wexler, K., & Redmond, S.M. (1999). Grammaticality judgments of an extended optional infinitive grammar: Evidence from English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 42*, 943-961.

Selzam, S., Krapohl, E., von Stumm, S., O'Reilly, P. F., Rimfeld, K., Kovas, Y., ... & Plomin, R. (2016). Predicting educational achievement from DNA. *Molecular Psychiatry*.

Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Rimfeld, K., Krapohl, E., Haworth, C. M. A., Dale, P. S., & Plomin, R. (2013). Strong genetic influence on an UK nationalwide test of educational achievement at the end of compulsory education at age 16. *PLoS ONE, 8(12).*

Spaulding, T.J., Plante, E., Farinella, K.A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37*, 61-72

Spinath, F. M., Price, T. S., Dale, P. S., & Plomin, R. (2004). The genetic and environmental origins of language disability and ability. *Child Development, 75*, 445-454.

Su, B. M., & Chan, D. K. (2017). Prevalence of hearing loss in US children and adolescents.. *JAMA Otolaryngology and Head and Neck Surgery.* Doi: 10.1001/jamaoto.2017.0953.

Wexler, K. 1996 The development of inflection in a biologically based theory of language acquisition. In M.L. Rice, *Toward a genetics of language*, Mahwah, NJ, Lawrence Erlbaum Associates.

Figure Legend

Figure 1. Distribution of GJ-20 scores in entire sample.

Table 1

Performance (as z-scores) on the Grammatical Judgment, Vocabulary, and Figurative Language tests overall, and by sex and zygosity

Measure (all age-corrected)	**Total M (SD) n**	**Female M (SD) n**	**Male M (SD) n**	**Monozygotic M (SD) n**	**Dizygotic M (SD) n**	**Sex F^1**	**Zygosity F^1**	**R^2**
GJ-20	.017 (.97) 405	.018 (.98) 204	.016 (.96) 201	.05 (1.01) 148	-.002 (.95) 257	0.01	0.27	0.00
Vocabulary	-.008 (.98) 151	-.005 (.94) 83	.01 (1.04) 68	-.23 (.99) 64	.16 (.94) 87	0.03	5.42*	0.03
Figurative Language	.04 (.97) 146	.03 (1.01) 80	.054 (.92) 66	-.19 (.96) 63	.22 (.94) 83	0.04	6.37*	0.03

Note. All computed values are based on one randomly selected twin from each pair. GJ-20 = A' measure of discrimination on test of finiteness marking; Vocabulary = Mill Hill Vocabulary Scale; Figurative Language = subtest of the Test of Language Competence – Expanded Edition.

[1] *F* for ANOVA performed on normalized measure, using one randomly selected twin per pair, to test main effects of sex and zygosity. * $p < .05$; ** $p < .01$. R^2 = proportion of variance explained by sex, zygosity, and their interaction

Table 2

Correlations among language measures and maternal education

Measure	Age-adjusted GJ-20 N	Age-Adjusted Vocabulary N	Age-adjusted Figurative Lang. N
Age-adjusted Vocabulary	0.12 151		
Age adjusted Figurative Lang.	.24 ** 146	.4 ** 145	
Maternal Education at first contact	.16 ** 395	.22 ** 148	.24 ** 143

Note. ** $p < .01$

Note. All computed values are based on one randomly selected twin from each pair. Vocabulary = Mill Hill Vocabulary Scale; Figurative Language = subtest of the Test of Language Competence – Expanded Edition; Maternal Education obtained at entry to study when twins were 18 months of age.

Table 3

Relation of low score on grammatical judgment task to other measures

Measure	**Mean (SD) or % for individuals with age-adjusted GJ-20 in the lowest 10% (N)**	**Mean (SD) or % for individuals with normal range age-adjusted GJ-20 (N)**	**Difference between low GJ-20 and normal-range GJ-20 (including only 1 randomly selected twin per pair)**	**Difference between low GJ-20 and normal-range GJ-20 (both twins included)**	**Difference with both twins included; controlling for non-independence of data**
% male	53 (81)	49 (757)			
% family history of language/literacy difficulties	11.1 (81)	8 (757)	Chi-Square= .778 (df=1); p=.264	Chi-Square= 1.060 (df=1); p=.203	
% reported to be dyslexic	36.4 (33)	8 (684)	Chi-square =1.725 (df=1); p=.153	Chi-square =.367 (df=1); p=.345	
% reported to have learning difficulties	54 (63)	19 (684)	Chi-square=18.74 (df=1); p<.001	Chi-square=44.29 (df=1); p<.001	
Maternal education at first contact M (SD)	-.30 (1.00) (79)	0.03 (.99) (709)			Chi-square=7.11 (df=1), p=0.03
Age adjusted Age 16 Vocabulary M (SD)	-.68 (.76) (22)	.06 (.99) (274)			Chi-Square=8.99 (df=1), p=0.003
Age-adjusted Age 16 Figurative language M (SD)	-.66 (1.12) (22)	.05 (.97) (267)			Chi-square=5.27 (df=1), p=0.02
GCSE – English M (SD)	-.61 (1.02) (54)	.05 (.98) (640)			Chi-square=13.25 (df=1), p<.001

GCSE – Mathematics M (SD)	-.57 (1.27) (57)	.05 (.98) (640)	Chi-square=5.67 (df=1), p=0.017

Note. Family history = reported language and/or reading learning difficulties in first-degree relative; Reported dyslexia = parental report at child age 7 years; Reported learning difficulties = parental report at child age 7 year; Vocabulary = Mill Hill Vocabulary Scale; Figurative Language = subtest of the Test of Language Competence – Expanded Edition; Maternal Education obtained at entry to study when twins were 18 months of age; GCSE-English = Composite of English language and English literature performance on the General Certificate of Secondary Education examination; GCSE – Mathematics = Mathematics performance on the General Certificate of Secondary Education.

Table 4

Probandwise concordance for placement in the low score category in GJ-20 measure

Criterion	N probands	MZ concordance	N MZ probands	DZ concordance	N DZ probands
Lowest 10%	81	.28	29	.19	52
Lowest 7%	54	.40	20	.24	34
Lowest 5%	40	.43	14	.15	26

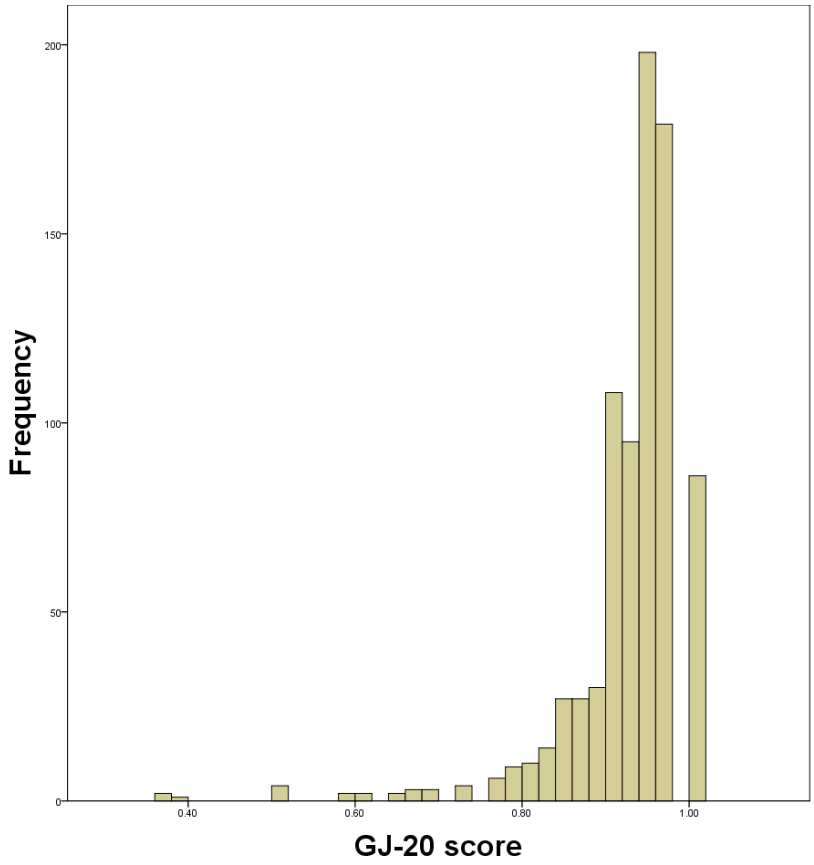
Note. Concordance calculated as the proportion of probands (individuals whose GJ-20 score place them in the low extreme category) whose co-twins are also affected.

Table 5

Liability threshold modeling of grammatical judgment (GJ-20) scores at four criteria for low performance

Criterion (n)	A	C	E
Lowest 10% (81)	.36 (0+ - .64)	.07 (0+ - .45)	.58 (.36 - .94)
Lowest 7% (54)	.47 (0+ - .84)	.20 (0+ - .64)	.33 (.16 - .59)
Lowest 5% (40)	.74 (0+-.93)	0 (0+-.67)	.26 (.07-.76)

Note. A, C, and E measure the degree of influence of genetic, shared environment, and nonshared environmental factors, respectively, on placement in the low category on this measure. A + C + E = 1.0. The notation 0+ indicates a value < .001, but still above zero. 95% confidence intervals are provided in parentheses.


Frequency
200
150
100
50
0
0.40
0.60
0.80
1.00
GJ-20 score